New Hampshire Statewide Assessment System

2018–2019

Volume 4
Evidence of Reliability and Validity

New Hampshire
Department of Education
Serving New Hampshire's Education Community

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

1. Introduction and Overview of Reliability and Validity Evidence

The state of New Hampshire implemented a new assessment program for operational use beginning in the 2017–2018 school year. This new program, named the New Hampshire Statewide Assessment System (NH SAS), replaced the Smarter Balanced Assessment Consortium (SBAC) in English language arts (ELA) and mathematics, and the New England Common Assessment Program (NECAP) in science. It is delivered as an online, computer-adaptive test (CAT) for ELA and mathematics and as an online, linear-on-the-fly test (LOFT) for science. The accommodation versions are generally available for students for whom there is a documented need on an Individualized Education Plan (IEP) or Section 504 Plan. Table 1 displays the complete list of test administration methods for the NH SAS.

Table 1: Test Administration

Subject	Administration	Grade
ELA Reading	Online Adaptive	3–8
ELA Writing	Online	3–8
Mathematics	Online Adaptive	3–8
Science	Online Linear-on-the-Fly	5, 8, 11

With the implementation of these tests, both reliability evidence and validity evidence are necessary to support appropriate inferences of student academic performance from the NH SAS scores. This volume provides empirical evidence about the reliability and validity of the 2018–2019 NH SAS, given its intended uses.

The purpose of this volume is to provide empirical evidence to support the following:

- **Reliability**. The reliability estimates are presented by grade and subject. This section also includes conditional standard errors of measurement (CSEM) and classification accuracy and consistency results by grade and subject.
- **Content validity**. Evidence is provided to show that test forms were constructed to measure the New Hampshire College and Career Ready Standards (NH CCRS) with a sufficient number of items targeting each area of the blueprint.
- **Internal structure validity**. Evidence is provided regarding the internal relationships among the subscale scores to support their use and to justify the item response theory (IRT) measurement model. This type of evidence includes observed and disattenuated Pearson correlations among reporting categories per grade.
- **Relationship of test scores to external variables**. Evidence of convergent and discriminant validity is provided using observed and disattenuated subscore correlations both within and across subjects. The correlations between interim and summative assessments, as well as the correlation between SBAC spring 2017 and NH SAS spring 2018 summative assessments in ELA and mathematics, are also presented.

- **Test fairness**. Fairness is analyzed statistically using differential item functioning (DIF) in tandem with content alignment reviews by specialists.

1.1 RELIABILITY

Reliability refers to consistency in test scores. Reliability can be defined as the degree to which individuals' deviation scores remain relatively consistent over repeated administrations of the same test or alternate test forms (Crocker & Algina, 1986). For example, if a person takes the same or parallel tests repeatedly, he or she should receive consistent results. The reliability coefficient refers to the ratio of true score variance to observed score variance:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}.$$

Another way to view reliability is to consider its relationship with the standard errors of measurement (SEM); the smaller the standard error, the higher the precision of the test scores. For example, classical test theory (CTT) assumes that an observed score (X) of each individual can be expressed as a true score (T) plus some error (E), $X = T + E$. The variance of X can be shown to be the sum of two orthogonal variance components:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2.$$

Returning to the definition of reliability as the ratio of true score variance to observed score variance, we can arrive at the following:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}.$$

Unlike the CTT, SEM in IRT varies over the ability continuum. These heterogeneous errors are a function of a test information function (TIF) that provides different information about test takers depending on their estimated abilities. Often, the TIF is maximized over an important performance cut, such as the *Proficient* cut score.

Because the TIF indicates the amount of information provided by the test at different points along the ability scale, its inverse indicates the "lack" of information at different points along the ability scale. This lack of information is the uncertainty, or the measurement error, of the score at various score points. Conventionally, fixed-form tests are maximized near the middle of the score distribution, or near an important classification cut, and have less information at the tails of the score distribution. See Section 3.3 for the derivation of heterogeneous errors in IRT.

1.2 VALIDITY

Validity refers to the degree to which "evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014). Messick (1989) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment." Both of these definitions emphasize evidence and theory to support inferences and interpretations of test scores. *The Standards* (AERA, APA, & NCME, 2014) suggests five sources of validity evidence that can be used in evaluating a proposed interpretation of test scores. When validating test scores, these sources of evidence should be considered carefully.

The first source of evidence for validity is the relationship between the test content and the intended test construct (see Section 4). In order for test score inferences to support a validity claim, the items should be representative of the content domain, and the content domain should be relevant to the proposed interpretation of test scores. To determine content representativeness, diverse panels of content experts conduct alignment studies, in which experts review individual items and rate them based on how well they match the test specifications or cognitive skills required for a particular construct (see Volume 2 for details). Evidence based on test content is a crucial component of validity; construct underrepresentation or irrelevancy could result in unfair advantages or disadvantages to one or more group of test takers.

Additionally, technology-enhanced items should be examined to ensure that no construct-irrelevant variance is introduced. If some aspect of the technology impedes, or advantages, a student in his or her responses to items, this could affect item responses and inferences regarding abilities on the measured construct (see Volume 2, Section 2.1).

The second source of validity evidence is based on "the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (AERA, APA, & NCME, 2014). This evidence is collected by surveying test takers about their performance strategies or responses to particular items. Because items are developed to measure particular constructs and intellectual processes, evidence that test takers have engaged in relevant performance strategies to correctly answer the items then supports the validity of the test scores.

The third source of evidence for validity is based on internal structure: the degree to which the relationships among test items and test components relate to the construct on which the proposed test scores are interpreted. DIF, which determines whether particular items may function differently for subgroups of test takers, is one method for analyzing the internal structure of tests (see Volume 1, Section 4.5). Other possible analyses to examine internal structure are dimensionality assessment, goodness-of-model-fit to data, and reliability analysis (see Sections 3 and 5 for details).

A fourth source of evidence for validity is the relationship of test scores to external variables. *The Standards* (AERA, APA, & NCME, 2014) divides this source of evidence into three parts: convergent and discriminant evidence, test-criterion relationships, and validity generalization. Convergent evidence supports the relationship between the test and other measures intended to assess similar constructs. Conversely, discriminant evidence differentiates the test from other measures intended to assess different constructs. To analyze both convergent and discriminant evidence, a multi-trait-multimethod matrix can be used. Additionally, test-criterion relationships indicate how accurately test scores predict criterion performance. The degree of accuracy mainly depends upon the purpose of the test, such as classification, diagnosis, or selection. Test-criterion evidence is also used to investigate predictions of favoring different groups. Due to construct underrepresentation or construct-irrelevant components, the relation of test scores to a relevant criterion may differ from one group to another. Furthermore, validity generalization is related to whether the evidence is situation-specific or can be generalized across different settings and times. For example, sampling errors or range restriction may need to be considered to determine whether the conclusions of a test can be assumed for the larger population.

Fifth, the intended and unintended consequences of test use should be included in the test validation process. Determining the validity of the test should depend upon evidence directly related to the test; this process should not be influenced by external factors. For example, if an employer administers a test to determine hiring rates for different groups of people, an unequal distribution of skills related to the measurement construct does not necessarily imply a lack of validity for the test. However, if the unequal distribution of scores is in fact due to an unintended, confounding aspect of the test, this would interfere with the test's validity. As described in Volume 1 and further in this volume, test use should align with the intended purpose of the test.

Supporting a validity argument requires multiple sources of validity evidence. This then allows one to evaluate if sufficient evidence has been presented to support the intended uses and interpretations of the test scores. Thus, determining the validity of a test first requires an explicit statement regarding the intended uses of the test scores, and subsequently, evidence that the scores can be used to support these inferences.

2. Purpose of the New Hampshire Statewide Assessment System

The primary purpose of New Hampshire Statewide Assessment System (NH SAS) is to yield test scores at the student level and at other levels of aggregation that reflect student performance relative to the New Hampshire College and Career Ready Standards (NH CCRS). As opposed to norm-referenced tests that are designed to compare or rank all students with one another, the NH SAS is a criterion-referenced test that is designed to measure student performance on the NH CCRS in English language arts (ELA), mathematics, and science. The NH SAS standards and test blueprints are discussed in Volume 2, Test Development. The test was developed using the principles of evidence-centered design and adherence to the principles of universal design to ensure that all students have access to the test content. NH SAS results can also provide data for state and federal accountability systems.

NH SAS enhances teaching and student learning by measuring growth in student performance and providing immediate feedback to educators and parents that can be used to form instructional strategies to remediate or enrich instruction. Assessments can be used as an indicator to determine whether students in New Hampshire are ready with the knowledge and skills that are essential for college and career readiness. Test scores provide the information needed to evaluate students' learning progress and to implement strategies that can help teachers improve their instruction.

Volume 2, Test Development, describes in more detail about the NH SAS, NH CCRS, and test blueprints. This volume provides evidence of content validity in Section 4. The NH SAS test scores are a useful indicator for understanding individual students' academic performance on the New Hampshire standards and whether students are progressing in their performance over time. Additionally, both individual and aggregated scores can be used for measuring the reliability of the test. The reliability of the test scores can be found in Section 3 of this volume.

3. Reliability

3.1 Reliability for ELA and Mathematics

The New Hampshire Statewide Assessment System (NH SAS) ELA and mathematics tests are computer-adaptive testing (CAT) administrations. Because there is no set form in adaptive testing, marginal reliability was computed for the scale scores, taking into account the varying measurement errors across ability range. Marginal reliability is a measure of the overall reliability of an assessment based on the average conditional SEM, estimated at different points on the ability scale, for all students.

Marginal reliability ($\overline{\rho}$) is defined as

$$\overline{\rho} = [\sigma^2 - \left(\frac{\sum_{i=1}^{N} CSEM_i^2}{N}\right)]/\sigma^2,$$

where N is the number of students; $CSEM_i$ is the conditional SEM of the scale score for student i; and σ^2 is the variance of the scale score. The higher the reliability coefficient, the greater the precision of the test.

Table 2 presents the reliability coefficients for students in ELA and mathematics. The reliability coefficients for both subjects and all grades range from 0.89 to 0.93.

Table 2: Reliability Coefficients, ELA and Mathematics

Subject	Grade	Reliability	Subject	Grade	Reliability
ELA	3	0.89	Mathematics	3	0.92
	4	0.89		4	0.92
	5	0.89		5	0.91
	6	0.90		6	0.91
	7	0.90		7	0.90
	8	0.90		8	0.93

3.2 Reliability of Science

The reliability of science is computed in a similar way as the marginal reliability defined in Section 3.1, except that $CSEM_i$ is the conditional SEM of the overall ability estimate for student i; and σ^2 is the variance of the overall ability estimates. The marginal reliability of science for the overall sample is reported by grade in Table 3. The overall reliability ranges from 0.84 to 0.85. The reliability for students who received a complete test (18 items) is about the same as the overall reliability for both grades. Due to the new structure of the science test, AIR has also explored the relationship between reliability and other important factors such as the effect of nuisance dimension (see Volume 1, Section 5.2.1). It was found that if the local dependencies among assertions pertaining to the same item are ignored, the marginal reliability increases to approximately 0.90. Ignoring local dependencies could be achieved by either computing the maximum likelihood estimate (MLE) of ability under the unidimensional Rasch model or by setting the variance parameters to zero for all item clusters when computing the marginal maximum likelihood estimation (MMLE) of ability under the one-parameter logistic (1PL) bifactor model (see Volume 1, Section 6.2.1).

Table 3: Marginal Reliability Coefficients for Science

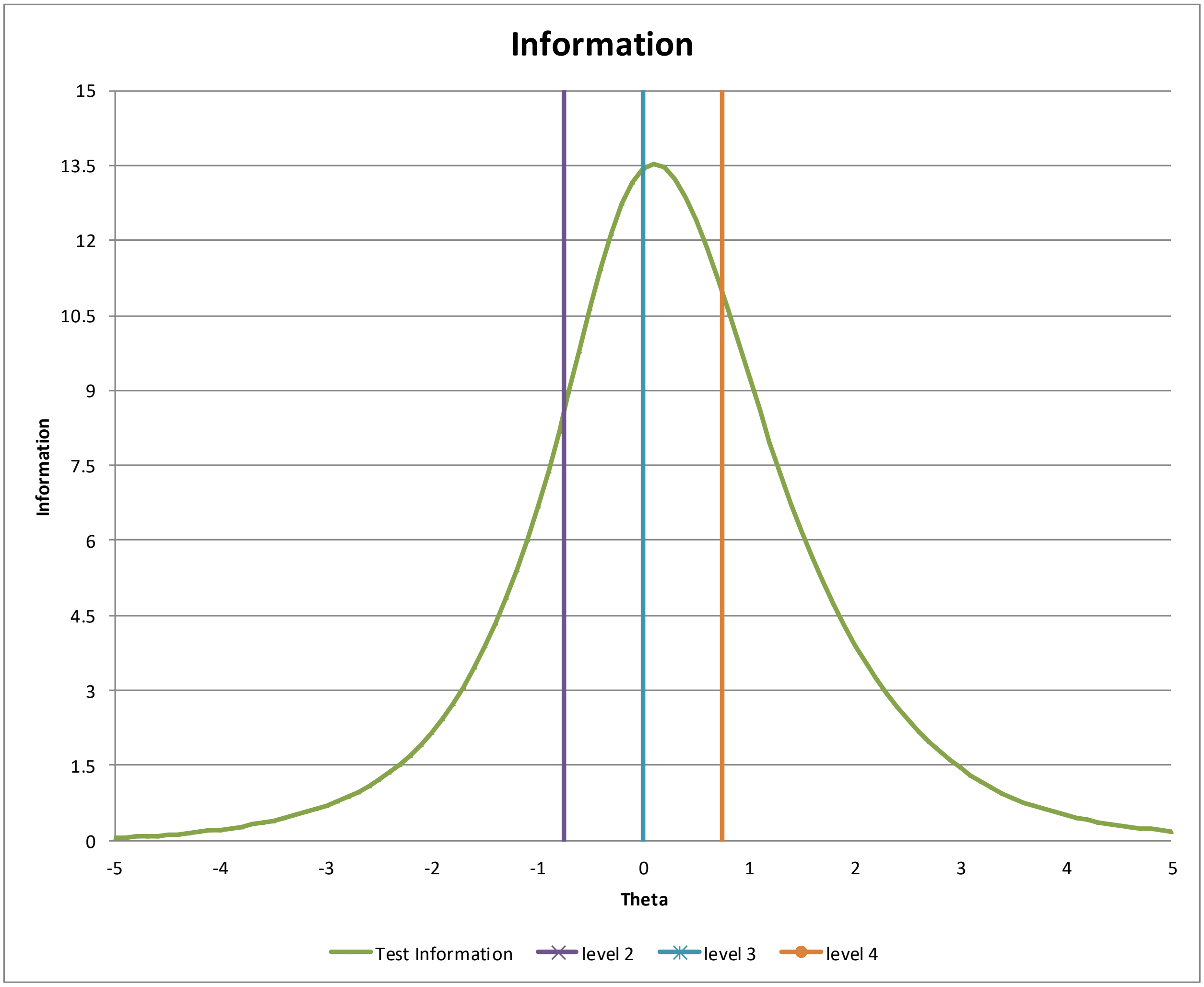
Grade	Sample Size	Reliability
5	13,187	0.84
8	12,060	0.84
11	11,385	0.85

3.3 Test Information Curves and Standard Error of Measurement for ELA and Mathematics

Within the IRT framework, measurement error varies across the range of ability as a result of the TIF. The TIF describes the amount of information provided by the test at each score point along the ability continuum. The inverse of the TIF is characterized as the conditional measurement error at each score point. For instance, if the measurement error is large, then less information is being provided by the assessment at that specific ability level.

Figure 1 displays a sample with three vertical lines indicating the performance cuts. The graphic shows that this test information is maximized in the middle of the score distribution, meaning that it provides the most precise scores in this range. Where the curve is lower at the tails indicates that the test provides less information about the test takers, relative to the center.

Figure 1: Sample Test Information Function



Computing these TIFs is useful in evaluating where the test is maximally informative. In IRT, the TIF is based on the estimates of the item parameters in the test, and the formula used for the NH SAS is calculated as:

$$TIF(\theta_s) = \sum_{i=1}^{N_{GPCM}} D^2 a_i^2 \left(\frac{\sum_{j=1}^{m_i} j^2 Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)} - \left(\frac{\sum_{j=1}^{m_i} jExp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)}{1 + \sum_{j=1}^{m_i} Exp\left(\sum_{k=1}^{j} Da_i(\theta_s - b_{ik})\right)} \right)^2 \right) + \sum_{i=1}^{N_{3PL}} D^2 a_i^2 \left(\frac{Q_i}{P_i} \left[\frac{P_i - c_i}{1 - c_i} \right]^2 \right),$$

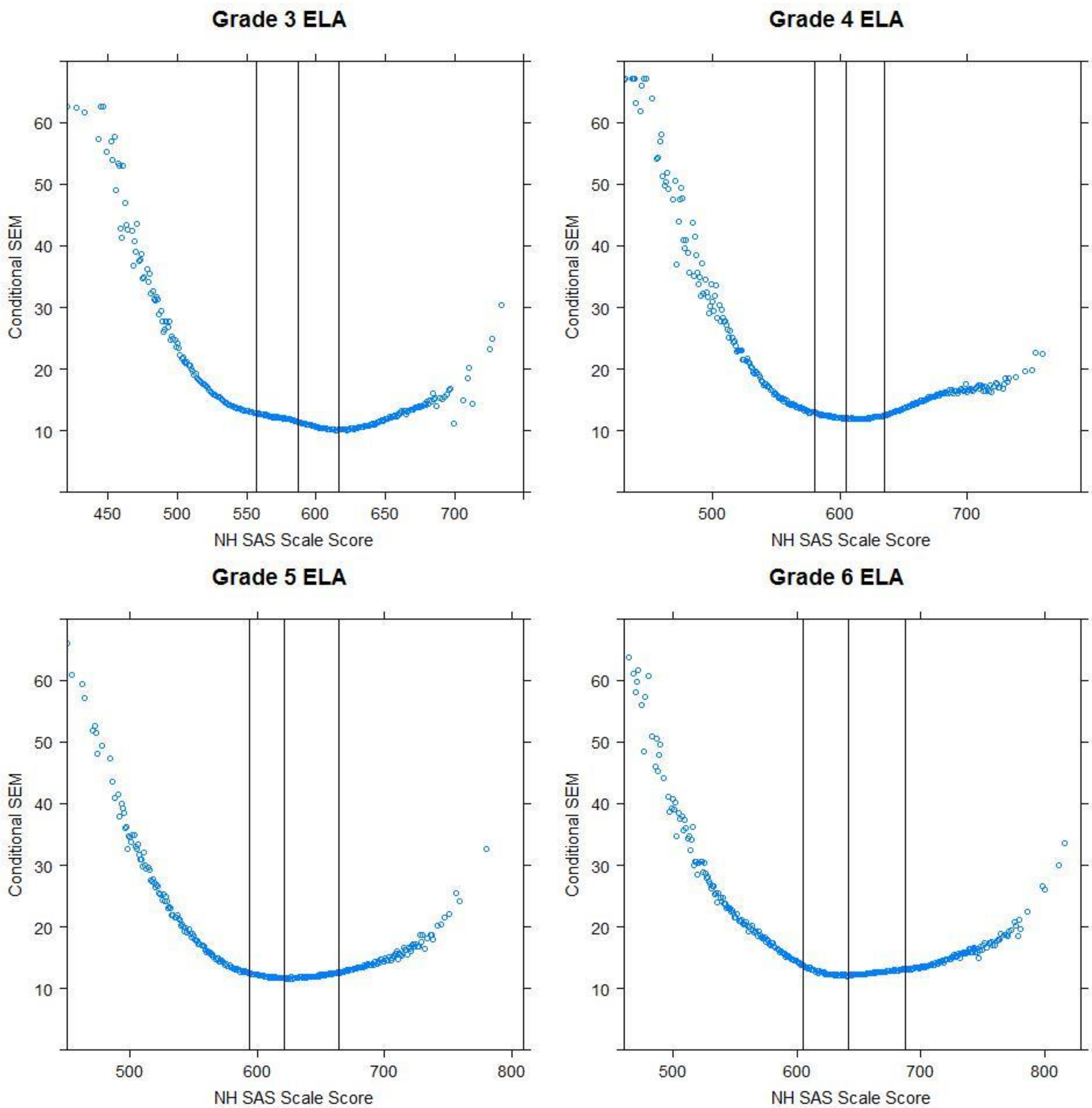
where N_{GPCM} is the number of items that are scored using the generalized partial credit model (GPCM) items; N_{3PL} is the number of items scored using 3PL or 2PL model; *i* indicates item *i* ($i \in \{1,2,\ldots,N\}$); m_i is the maximum possible score of the item; *s* indicates student *s*; and θ_s is the ability of student *s*.

The standard error for estimated student ability (theta score) is the square root of the reciprocal of the TIF:

$$se(\theta) = \frac{1}{\sqrt{TIF(\theta_i)}}.$$

It is typically more useful to consider the inverse of the TIF rather than the TIF itself, as the standard errors are more useful for score interpretation. For this reason, standard error plots are presented in Figure 2 and Figure 3, respectively, instead of the TIFs for ELA and mathematics. The plots presented in this section are based on the scaled scores reported in spring 2019. Vertical lines represent the three performance-level cut scores.

Figure 2: Conditional Standard Errors of Measurement for ELA



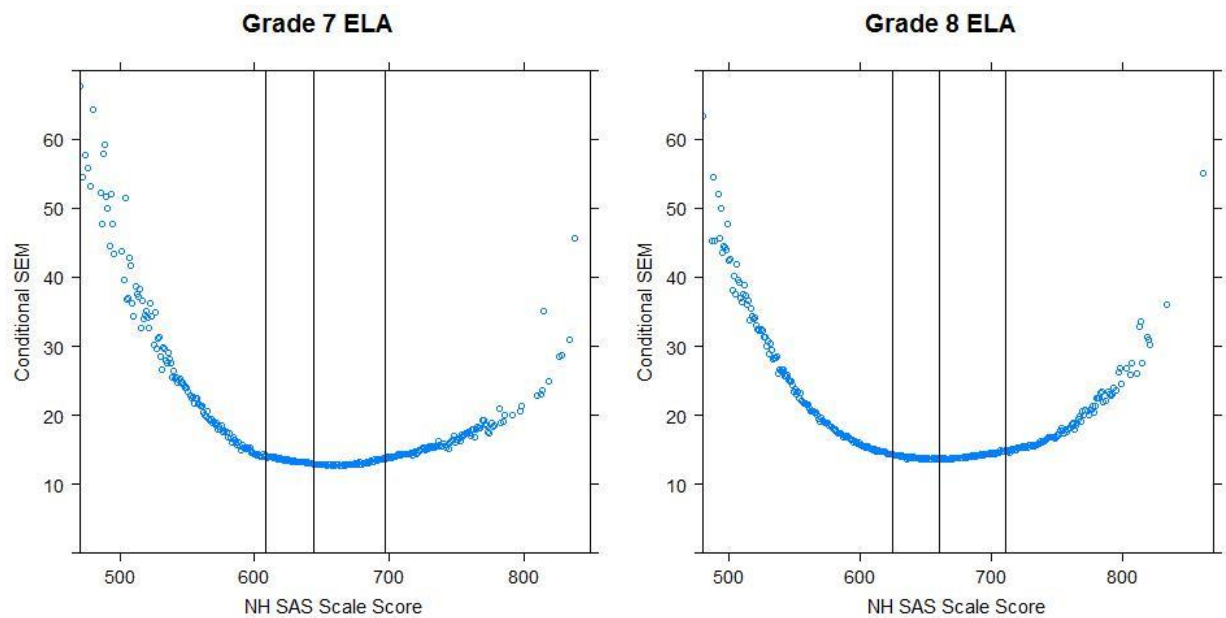
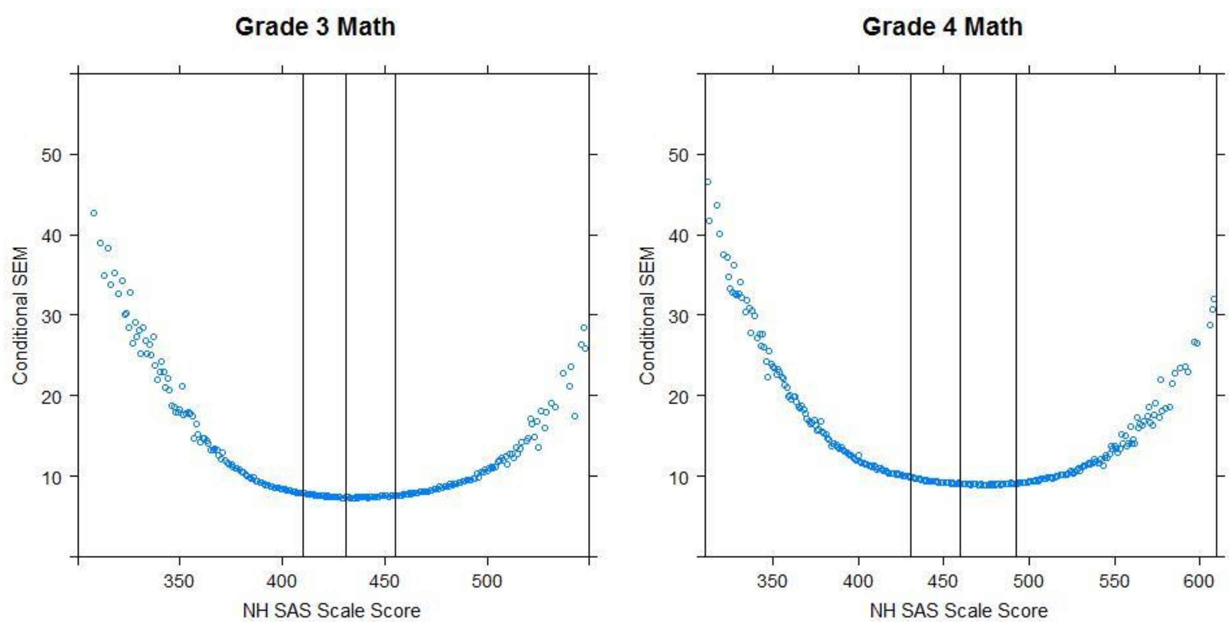
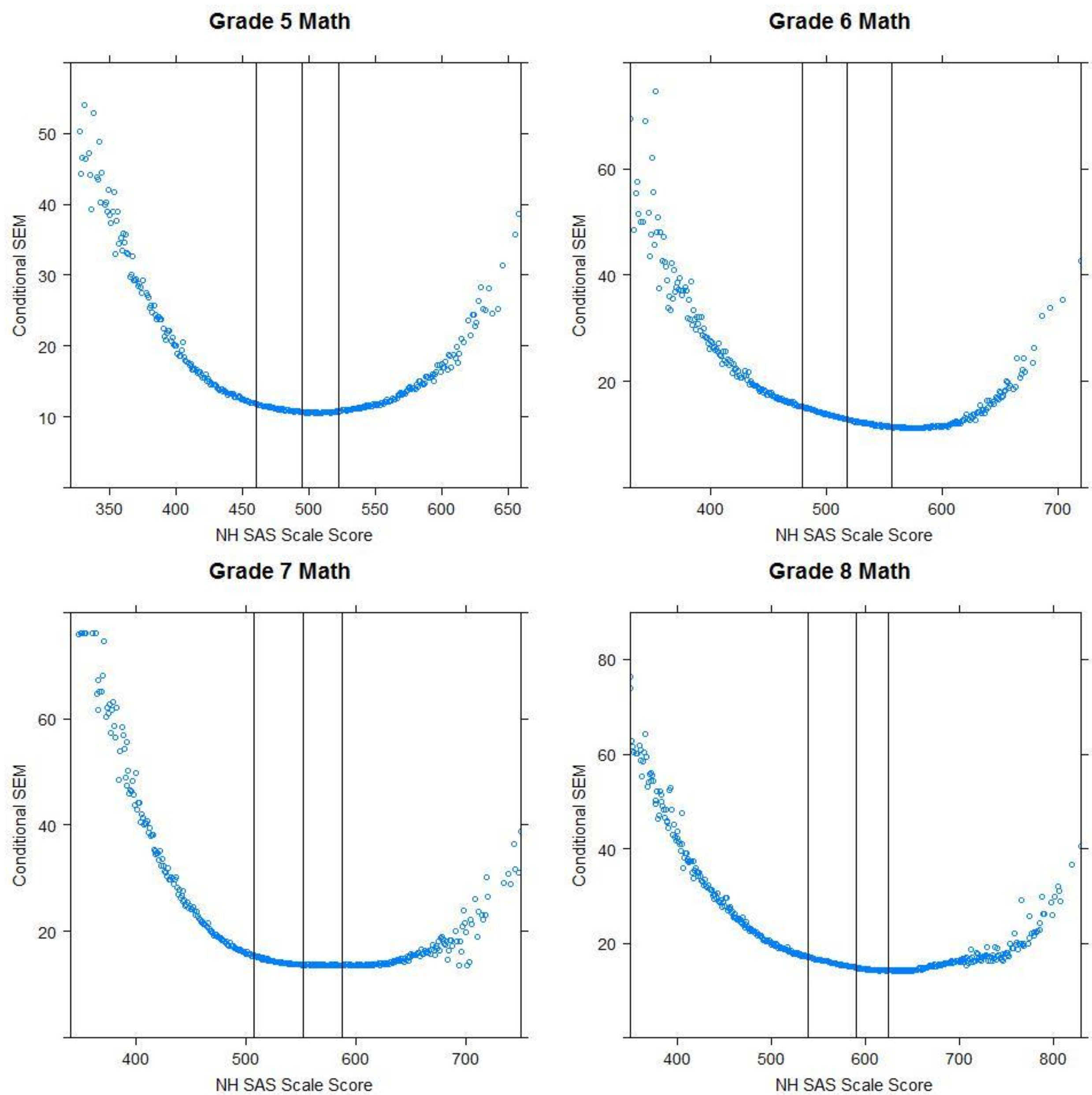


Figure 3: Conditional Standard Errors of Measurement for Mathematics




Grade 5 Math
Conditional SEM
NH SAS Scale Score
Grade 6 Math
Conditional SEM
NH SAS Scale Score
Grade 7 Math
Conditional SEM
NH SAS Scale Score
Grade 8 Math
Conditional SEM
NH SAS Scale Score

For most tests, the standard error curves follow the typical expected trends with more test information regarding scores observed near the middle of the score scale. In some grades in ELA and mathematics, the highest test information is observed at the *Proficient* and *Above Proficient* performance-level cuts.

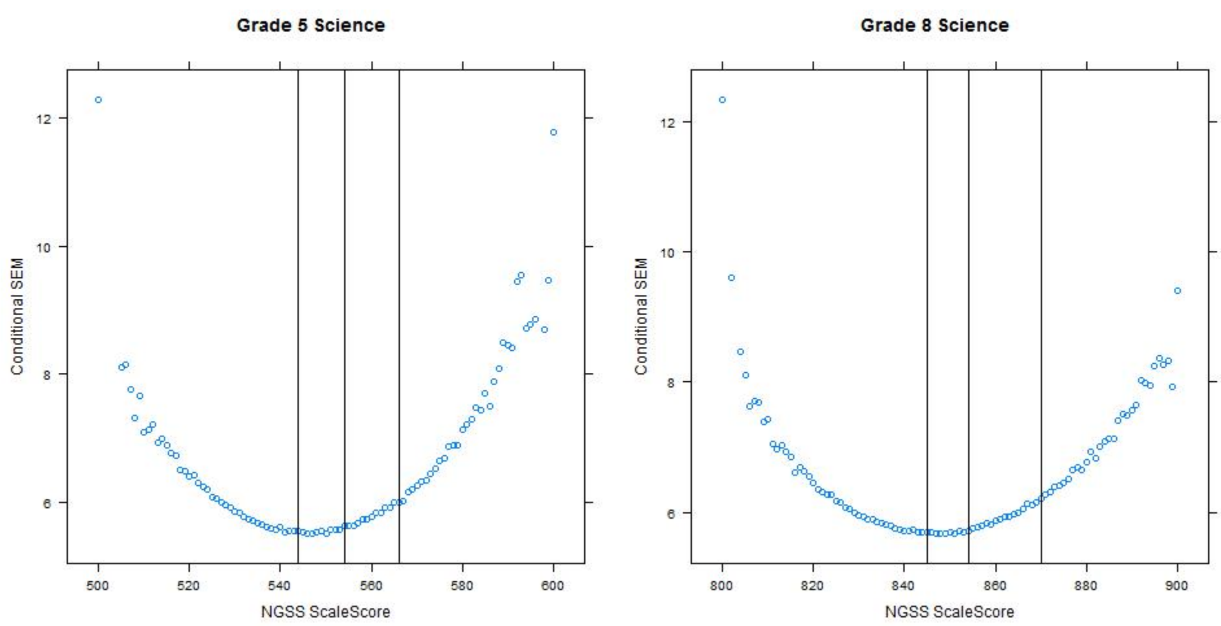
Overall, the standard error curves suggest that students are measured with a high degree of precision, given that the standard errors are consistently low. However, larger standard errors are observed at the lower ends of the score distribution relative to the higher ends. This occurs because the item pools currently have a shortage of easy items that are better targeted toward these lower achieving students. Content experts use this information to consider how to further target and populate item pools.

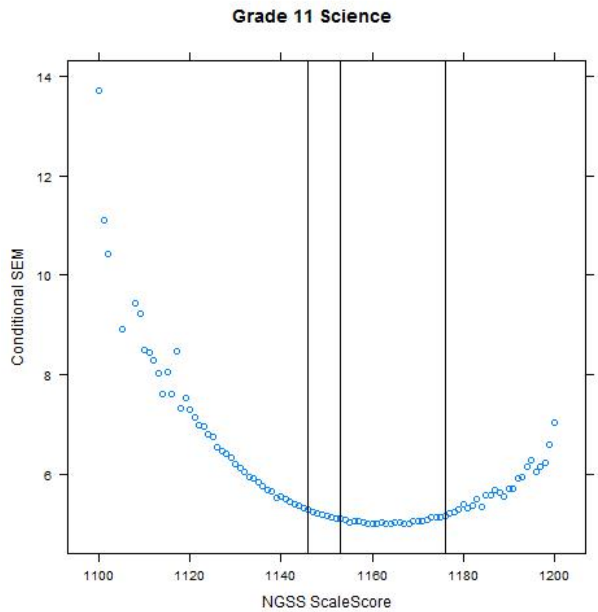
Appendix B includes scale score by scale score CSEM and corresponding performance levels for each scale score. The SEM for each reporting category is also presented in Appendix A.

3.4 Standard Error of Measurement for Science

The computation method of conditional standard error for science has been described in Section 6.2 of Volume 1. Figure 4 presents the conditional standard error curves for science. The lowest standard errors are observed near the proficiency cut scores for both grades, which is a desirable test property.

Figure 4: Conditional Standard Errors of Measurement for Science





3.5 Reliability of Performance Classification

When student achievement is reported in terms of performance levels, a reliability of classifying students into a specific level can be computed in terms of the likelihood of accurate and consistent classification as specified in Standard 2.16 in *The Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014).

The reliability of performance classification can be examined in terms of classification accuracy and classification consistency. Classification accuracy refers to the agreement between the classifications based on the form actually taken and the classifications that would be made based on the students' true scores, if they could hypothetically be obtained. Classification consistency refers to the agreement between the classifications based on the form actually taken and the classifications that would be made on the basis of an alternate, equivalently constructed test form.

In reality, the true ability is unknown, and students are not administered an alternate, equivalent form. Therefore, classification accuracy and consistency are estimated based on students' item scores, the item parameters, and the assumed latent ability distribution as described in the following sections. The true score is an expected value of the test score with measurement error.

For student j, the student's estimated ability is $\hat{\theta}_j$ with SEM of $se(\hat{\theta}_j)$, and the estimated ability is distributed as $\hat{\theta}_j \sim N\left(\theta_j, se^2(\hat{\theta}_j)\right)$, assuming a normal distribution, where θ_j is the unknown true ability of student j. The probability of the true score at performance level l $(l = 1, \cdots, L)$ is estimated as

$$p_{jl} = p(c_{Ll} \le \theta_i < c_{Ul}) = p\left(\frac{c_{Ll} - \hat{\theta}_j}{se(\hat{\theta}_j)} \le \frac{\theta_j - \hat{\theta}_j}{se(\hat{\theta}_j)} < \frac{c_{Ul} - \hat{\theta}_j}{se(\hat{\theta}_j)}\right)$$
$$= p\left(\frac{\hat{\theta}_j - c_{Ul}}{se(\hat{\theta}_j)} < \frac{\hat{\theta}_j - \theta_j}{se(\hat{\theta}_j)} \le \frac{\hat{\theta}_j - c_{Ll}}{se(\hat{\theta}_j)}\right) = \Phi\left(\frac{\hat{\theta}_j - c_{Ll}}{se(\hat{\theta}_j)}\right) - \Phi\left(\frac{\hat{\theta}_j - c_{Ul}}{se(\hat{\theta}_j)}\right),$$

where c_{Ll} and c_{Ul} denote the score corresponding to the lower and upper limits of the performance level l, respectively.

3.5.1 Classification Accuracy

Using p_{jl}, the expected number of students at level l, based on students from observed level *k*, can be expressed as

$$E_{Akl} = \sum_{pl_j \in k} p_{jl},$$

where pl_j is the jth student's performance level, the values of E_{Akl} are the elements used to populate the matrix $\boldsymbol{E_A}$, a $L \times L$ matrix of conditionally expected numbers of students to score within each performance level, based on their true scores. The classification accuracy (*CA*) at level l is estimated by

$$CA_l = \frac{E_{Akl}}{N_k},$$

where N_k is the observed number of students scoring in performance level k.

The classification accuracy for the pth cut is estimated by forming square, partitioned blocks of the matrix $\boldsymbol{E_A}$ and summing all the elements within the block as follows:

$$CAC = \left(\sum_{k=1}^{p} \sum_{l=1}^{p} E_{Akl} + \sum_{k=p+1}^{L} \sum_{l=p+1}^{L} E_{Akl} \right) \Big/ N,$$

where N is the total number of students.

The overall classification accuracy is estimated from the diagonal elements of the matrix:

$$CA = \frac{tr(\boldsymbol{E_A})}{N}.$$

Table 4 through Table 6 provide the overall classification accuracy and the classification accuracy for the individual cuts for ELA, mathematics, and science, respectively. The overall classification accuracy of the tests ranges from 76% to 80% for ELA, from 79% to 80% for mathematics, and from 72% to 77% for science. The cut accuracy rates are high across all grades and subjects with a minimum value of 91% for ELA, 92% for mathematics, and 89% for science. This denotes that more than 88% of the time, we can accurately differentiate students between adjacent performance levels in the spring 2019 NH SAS.

Table 4: Classification Accuracy Index, ELA

Grade	Overall Accuracy (%)	Cut Accuracy (%)		
		Cut 1	Cut 2	Cut 3
3	77.26	93.00	91.52	92.65

Grade	Overall Accuracy (%)	Cut Accuracy (%)		
		Cut 1	Cut 2	Cut 3
4	76.15	92.68	91.09	92.13
5	77.85	93.61	91.43	92.71
6	78.97	94.38	91.38	93.20
7	79.63	94.09	91.25	94.27
8	78.90	93.98	91.13	93.75

Table 5: Classification Accuracy Index, Mathematics

Grade	Overall Accuracy (%)	Cut Accuracy (%)		
		Cut 1	Cut 2	Cut 3
3	79.89	94.52	92.14	93.21
4	80.16	94.15	91.74	94.26
5	79.38	92.80	92.21	94.33
6	79.21	93.04	91.51	94.63
7	79.39	93.72	91.63	93.99
8	80.43	93.58	93.03	93.76

Table 6: Classification Accuracy Index, Science

Grade	Overall Accuracy (%)	Cut Accuracy (%)		
		Cut 1	Cut 2	Cut 3
5	71.79	89.37	88.82	93.02
8	73.67	89.09	89.15	94.72
11	77.10	88.67	89.75	97.55

3.5.2 Classification Consistency

Assuming the test is administered twice independently to the same group of students, similarly to accuracy, a $L \times L$ matrix $\boldsymbol{E_C}$ can be constructed. The element of $\boldsymbol{E_C}$ is populated by

$$E_{Ckl} = \sum_{j=1}^{N} p_{jl} p_{jk},$$

where p_{jl} is the probability of the true score at performance level l in test one, and p_{jk} is the probability of the true score at performance level k in test two for the jth student. The classification consistency index for the cuts (*CCC*) and overall classification consistency (*CC*) were estimated in a way similar to CAC and CA.

$$CCC = \left(\sum_{k=1}^{p} \sum_{l=1}^{p} E_{Ckl} + \sum_{k=p+1}^{L} \sum_{l=p+1}^{L} E_{Ckl} \right) \Big/ N,$$

and

$$CC = \frac{tr(\boldsymbol{E_C})}{N}.$$

Table 7 through Table 9 provide the classification consistency, both overall and of the individual cuts for ELA, mathematics, and science, respectively. The overall classification consistency of the test ranges from 68% to 72% for ELA, from 71% to 73% for mathematics, and from 63% to 70% for science.

The individual cut consistency rates are high across all grades and subjects, with the minimum values of 87% for ELA, 88% for mathematics, and 84% for science. In all performance levels, classification accuracy is higher than classification consistency. Classification consistency rates can be lower than classification accuracy; the consistency is based on two tests with measurement errors, but the accuracy is based on one test with a measurement error and the true score. The accuracy and consistency rates for each performance level are higher for the levels with smaller standard error.

Table 7: Classification Consistency Index, ELA

Grade	Overall Consistency (%)	Cut Consistency (%)		
		Cut 1	Cut 2	Cut 3
3	68.65	90.14	88.09	89.60
4	67.56	89.69	87.42	88.86
5	69.32	90.89	87.87	89.82
6	70.58	92.00	87.85	90.50
7	71.59	91.63	87.70	91.93
8	70.59	91.41	87.55	91.20

Table 8: Classification Consistency Index, Mathematics

Grade	Overall Consistency (%)	Cut Consistency (%)		
		Cut 1	Cut 2	Cut 3
3	72.03	92.25	88.97	90.45
4	72.35	91.80	88.42	91.93
5	71.34	89.85	88.99	91.99
6	71.09	90.21	88.04	92.41
7	71.34	91.08	88.24	91.47
8	73.02	90.99	90.16	91.24

Table 9: Classification Consistency Index, Science

Grade	Overall Consistency (%)	Cut Consistency (%)		
		Cut 1	Cut 2	Cut 3
5	62.75	85.10	84.37	90.22
8	65.11	84.74	84.83	92.57
11	69.74	84.12	85.59	96.47

3.6 Precision at Cut Scores

Table 10 through Table 12 present the mean CSEM at each performance level by grade and subject. These tables also include performance-level cut scores and associated CSEM.

Table 10: Performance Levels and Associated CSEM, ELA

Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
3	1	17.12	-	-
	2	12.18	557	12.82
	3	10.68	587	11.37
	4	10.97	616	10.19
4	1	17.74	-	-
	2	12.39	580	13.1
	3	12.11	605	12.15
	4	14.32	635	12.57

Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
5	1	16.52	-	-
	2	12.02	594	12.48
	3	12.01	621	11.75
	4	13.61	664	12.62
6	1	18.59	-	-
	2	12.58	605	13.85
	3	12.6	642	12.2
	4	14.00	688	13.23
7	1	19.13	-	-
	2	13.48	608	14.28
	3	13.00	644	12.89
	4	14.76	697	13.64
8	1	17.95	-	-
	2	13.86	625	14.4
	3	14.09	661	13.74
	4	16.14	711	14.96

Table 11: Performance Levels and Associated CSEM, Mathematics

Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
3	1	10.8	-	-
	2	7.63	410	7.96
	3	7.49	431	7.47
	4	8.56	455	7.67
4	1	12.99	-	-
	2	9.41	431	9.91
	3	9.06	460	9.16
	4	10.19	492	9.13
5	1	16.55	-	-
	2	11.19	460	11.92
	3	10.68	495	10.74
	4	12.12	522	10.79

Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
6	1	21.49	-	-
	2	13.98	479	15.39
	3	12.15	518	12.94
	4	11.8	556	11.48
7	1	24.42	-	-
	2	14.35	507	15.58
	3	13.76	552	13.84
	4	14.32	587	13.67
8	1	24.17	-	-
	2	15.96	539	17.12
	3	14.49	591	14.94
	4	15.11	625	14.28

Table 12: Performance Levels and Associated CSEM, Science

Grade	Performance Level	Mean CSEM	Cut Score (Scale Score)	CSEM at Cut Score
5	1	6.52	-	-
	2	5.54	544	5.54
	3	5.77	554	5.62
	4	7.60	566	6.00
8	1	6.67	-	-
	2	5.69	845	5.69
	3	5.92	854	5.73
	4	7.28	870	6.22
11	1	7.18	-	-
	2	5.19	1146	5.29
	3	5.05	1153	5.10
	4	5.74	1176	5.17

3.7 ELA Writing Prompts Inter-Rater Reliability

Writing responses for the 2018–2019 school year were scored with AIR's AutoScoring Model. The validity of this machine-scoring system was assessed at the beginning of the testing window.

3.7.1 Automated Scoring Engine

AIR's essay scoring engine, AutoScore, uses a statistical process to evaluate writing prompts. Autoscore evaluates papers against the same rubric used by human raters, but a statistical process is used to analyze each paper and assign scores for each of the three dimensions. The engine uses the same process for scoring essays every time a new prompt is submitted.

Statistical rubrics are effectively proxy measures. Although they can directly measure some aspects of writing conventions (e.g., use of passive voice, misspellings, run-on sentences), they do not directly measure argument structure or content relevance. Hence, though statistical rubrics often prove useful for scoring essays and even for providing some diagnostic feedback in writing, they do not develop a sufficiently specific model of the correct semantic structure to score many propositional items. Furthermore, they cannot provide the explanatory or diagnostic information available from an explicit rubric. For example, the frequency of incorrect spellings may predict whether a response to a factual item is correct—higher-performing students may also have better spelling skills. Spelling may prove useful in predicting the human score, but it is not the actual reason that the human scorer deducts points. Indeed, statistical rubrics are not about explanation or reason but rather about a prediction of how a human would score the response.

AIR's essay-scoring engine uses a statistical rubric with great success, as measured by the rater agreements observed relative to the human-to-human rater agreements. This technology is similar to all essay-scoring systems in the field. Although some systems replace the statistical process with a "neural network" algorithm, that algorithm functions like the statistical model. Not all descriptions of essay-scoring algorithms are as transparent as AIR's, but whenever a training set is used for the machine to "learn a rubric," the same technology is being used.

The engine is designed to employ a "training set," a set of essays scored with maximally valid scores, that is used to form the basis of the prediction model. The quality of the human-assigned scores is critical to the identification of a valid model and the final performance of the scoring engine. Moreover, an ideal training sample over-represents higher- and lower-scoring papers and is selected according to a scientific sampling design with known probabilities of selection.

The training process of the scoring engine has two phases. The first phase requires oversampled, high- and low-scoring papers, leaving an equally weighted representative sample for the second phase. The first phase is used to identify concepts that are proportionately represented in higher-scoring papers. Here, concepts are defined as words and their synonyms, as well as clusters of words used meaningfully in proximity.

The second phase takes a series of measures on each essay in the remaining training set. These measures include latent semantic analysis (LSA) measures based on the concepts identified in the first phase; other semantic measures indicate the coherence of concepts within and across paragraphs and a range of word-use and syntactic measures. The LSA is similar to a data reduction method identifying common concepts within the narrative and reducing the data to a configurable number of LSA dimensions.

For each trait in the rubric, the system estimates an appropriate statistical model in which these LSA and other syntactic characteristics described earlier serve as the independent variables, and the final, resolved score serves as the dependent variable in an ordered probit regression. This model, along with its final parameter estimates, is used to generate a predicted or "proxy" score. The probability of scoring in the *p*th category is compared to a random draw from the uniform distribution, and a final score point of 1–4 is determined from this comparison.

In addition to the training set, an independent, random sample of responses is drawn for the cross-validation of the identified scoring rubric. As with the training set, student responses in the cross-validation study are handscored, and the LSA and other syntactic characteristics of the papers are computed. Subsequently, a second machine score is generated by applying the model coefficients obtained from the ordered probit in the training set. This forms a predicted score for the papers in the cross-validation set for each dimension in the rubric, which can then be used to evaluate the agreement rates between the human and Autoscore engine.

When implementing the scoring engine, we expect the computer-to-human agreement rates to be at least as high as the human-to-human agreement rates obtained from the double-scored process. If the engine yields scores with rater agreement rates that are at least as high as the human rater agreement rates, then the scoring engine can be deployed for operational scoring. If the computer-to-human agreement rates are not at least as high as the human-to-human rates, then adjustments to the scoring engine statistical model are necessary in order to find a scoring model that yields rater agreement rates that match the human-to-human rates.

To train AIR's artificial intelligence (AI) scoring engine, a subset of papers was selected using stratified random sampling and scored by two human raters. Essay responses to the AIRCore writing prompts were sent to the vendors Measurement Incorporated (MI) or Data Recognition Corporation (DRC) for human scoring. Using anchor papers selected by content experts and finalized rubrics (Table 13), human raters were trained to score writing responses at the rangefinding meeting. Raters revisited anchor papers and rubrics at rangefinding meetings to re-familiarize themselves with scoring, including a range of sample responses and scores.

At the rangefinding meeting, raters were assigned to groups. As training, the leader of each group read out loud student responses to raters; the raters independently referred back to the anchors and rubrics and they shared what they thought the score for the particular response should be. If the decision among raters was unanimous, they had a brief discussion and then moved to the next response. If the decision was not unanimous, the raters had a discussion referring to the anchors and rubrics to reach a consensus.

Table 13: Writing Rubrics

Dimension	Rubric	Maximum Score Point
Conventions	The response demonstrates an adequate command of basic conventions. The response may include the following: • Some minor errors in usage but no patterns of errors • Adequate use of punctuation, capitalization, sentence formation, and spelling	2
Evidence & Elaboration	The response provides thorough and convincing support, citing evidence for the controlling idea or main idea that includes the effective use of sources, facts, and details. The response includes most of the following: • Smoothly integrated, thorough, and relevant evidence, including precise references to sources • Effective use of a variety of elaborative techniques (including but not limited to definitions, quotations, and examples), demonstrating an understanding of the topic and text • Clear and effective expression of ideas, using precise language • Academic and domain-specific vocabulary clearly appropriate for the audience and purpose • Varied sentence structure, demonstrating language facility	4
Purpose, Focus, & Organization	The response is fully sustained and consistently focused within the purpose, audience, and task, and it has a clear controlling idea and effective organizational structure creating coherence and completeness. The response includes most of the following: • Strongly maintained controlling idea with little or no loosely related material • Skillful use of a variety of transitional strategies to clarify the relationships between and among ideas • Logical progression of ideas from beginning to end with a satisfying introduction and conclusion • Appropriate style and objective tone established and maintained	4

Two trained raters scored each writing item response. When scores from reader 1 and reader 2 were not in adjacent agreement, the response was sent for resolution scoring by a team leader or scoring director. The final item score was based on the resolution score, when present, or else on the initial read. Score discrepancies were resolved before being sent to AIR. Percentage agreement rates were computed to ensure that the machine scores are comparable to the human scores.

As seen in Table 14, exact agreement (when two raters gave the same score), adjacent rating (when the difference between two raters was 1), and non-adjacent rating (when the difference was larger than 1) were all determined. In this example, the exact agreement was 2/4, 50%, and the adjacent and non-adjacent percentages were 25% each.

Table 14: Rating Agreement Example

Response	Rater 1	Rater 2	Agreement
1	2	3	Adjacent
2	1	1	Exact

Response	Rater 1	Rater 2	Agreement
3	2	2	Exact
4	2	0	Non-Adjacent

Likewise, inter-rater reliability monitors how often scorers are in exact agreement with each other and ensures that an acceptable agreement rate is maintained. The calculations for inter-rater reliability in this report are as follows:

Percentage Exact: total number of responses by scorer in which scores are equal, divided by the number of responses that were scored twice

Percentage Adjacent: total number of responses by scorer in which scores are one score point apart, divided by the number of responses that were scored twice

Percentage Non-Adjacent: total number of responses by scorer in which scores are more than one score point apart, divided by the number of responses that were scored twice, when applicable

Table 15 displays percentage agreement in the training sample and validation sample. The total number of LSA dimensions and the sample size for validation are also presented in Table 15. In the training sample, the percentage of exact agreement ranged from 57% to 80%. The percentage of adjacent rating was between 19% and 41%. The non-adjacent percentages fell between 0% and 4%. In the validation sample, the percentage of exact agreement ranged from 68% to 84%. The percentage of adjacent rating was between 15% and 32%. The non-adjacent percentages fell between 0% and 2%. Table 15 shows that the scoring engine produced comparable results with human scores.

Table 15: Percentage Agreement in Handscoring and Scoring Engine

Grade	Item ID	Dimension	Handscoring in Training Sample				AIR Auto-Scoring Model in Validation Sample			
			% Exact	% Adjacent	% Non-Adjacent	LSA	% Exact	% Adjacent	% Non-Adjacent	N for comparison
3	7402	Conventions	63.56	34.89	1.56	40	71.78	26.44	1.78	450
		Evidence and Elaboration	63.21	33.49	3.30	10	70.99	28.77	0.24	424
		Purpose, Focus, and Organization	66.44	31.95	1.61	100	69.66	29.89	0.46	435
	7407	Conventions	69.84	29.71	0.45	40	75.51	23.81	0.68	441
		Evidence and Elaboration	56.82	40.68	2.50	40	67.50	32.05	0.45	440
		Purpose, Focus, and Organization	61.19	37.44	1.37	10	67.58	31.74	0.68	438
4	3084	Conventions	64.23	34.37	1.41	50	68.73	30.99	0.28	355
		Evidence and Elaboration	74.22	25.50	0.28	50	84.42	15.01	0.57	353
		Purpose, Focus, and Organization	70.51	28.95	0.54	10	79.36	20.38	0.27	373
	3086	Conventions	66.88	32.46	0.65	10	70.81	28.54	0.65	459
		Evidence and Elaboration	75.00	24.78	0.22	100	77.85	21.49	0.66	456
		Purpose, Focus, and Organization	72.03	27.09	0.88	10	76.21	23.57	0.22	454
5	3133	Conventions	70.92	28.66	0.42	40	75.31	24.48	0.21	478
		Evidence and Elaboration	69.47	30.32	0.21	100	78.95	20.84	0.21	475
		Purpose, Focus, and Organization	70.74	29.05	0.21	40	76.00	23.79	0.21	475
	4286	Conventions	74.95	24.20	0.86	50	73.66	26.34	0.00	467
		Evidence and Elaboration	64.10	34.14	1.76	100	75.11	24.45	0.44	454
		Purpose, Focus, and Organization	71.93	27.19	0.88	10	75.88	23.90	0.22	456

Grade	Item ID	Dimension	Handscoring in Training Sample				AIR Auto-Scoring Model in Validation Sample			
			% Exact	% Adjacent	% Non-Adjacent	LSA	% Exact	% Adjacent	% Non-Adjacent	N for comparison
6	3138	Conventions	67.99	31.57	0.44	50	73.95	26.05	0.00	453
		Evidence and Elaboration	64.85	33.41	1.75	40	75.33	24.45	0.22	458
		Purpose, Focus, and Organization	63.70	33.91	2.39	100	73.04	26.52	0.43	460
	5438	Conventions	67.02	31.50	1.48	50	76.53	23.26	0.21	473
		Evidence and Elaboration	65.34	30.48	4.18	40	71.82	27.97	0.21	479
		Purpose, Focus, and Organization	60.38	37.11	2.52	100	68.55	31.45	0.00	477
7	3037	Conventions	70.58	28.98	0.44	50	76.55	23.23	0.22	452
		Evidence and Elaboration	67.03	31.90	1.08	40	78.23	21.55	0.22	464
		Purpose, Focus, and Organization	65.56	33.33	1.10	10	77.92	22.08	0.00	453
	3883	Conventions	76.22	23.78	0.00	40	80.28	19.72	0.00	492
		Evidence and Elaboration	71.05	27.93	1.03	10	80.90	19.10	0.00	487
		Purpose, Focus, and Organization	68.10	30.88	1.02	50	78.53	21.27	0.20	489
8	3056	Conventions	77.90	21.44	0.66	10	80.31	19.47	0.22	457
		Evidence and Elaboration	75.89	23.44	0.67	40	75.22	24.33	0.45	448
		Purpose, Focus, and Organization	80.22	19.14	0.65	50	72.47	26.67	0.86	465
	3058	Conventions	78.01	21.76	0.23	40	83.56	16.44	0.00	432
		Evidence and Elaboration	74.43	25.11	0.45	50	75.11	24.66	0.23	442
		Purpose, Focus, and Organization	69.34	29.06	1.60	100	74.60	25.40	0.00	437

In addition to the percentage agreement rates, the quadratic-weighted kappa values were computed for the training sample and the validation sample for the writing prompts adopted in the spring 2019 NH SAS.

Cohen's kappa (Cohen, 1968) is an index of inter-rater agreement after accounting for the agreement that could be expected due to chance. This statistic can be computed as

$$K = \frac{P_o - P_c}{1 - P_c},$$

where P_o is the proportion of observed agreement, and P_c indicates the proportion of agreement by chance. Cohen's kappa treats all disagreement values with equal weights. Weighted kappa coefficients (Cohen, 1968), however, allow unequal weights, which can be used as a measure of validity. Weighted kappa coefficients were calculated using the formula below:

$$K_w = \frac{P'_o - P'_c}{1 - P'_c},$$

where

$$P'_o = \frac{\sum w_{ij} p_{oij}}{w_{max}},$$

$$P'_c = \frac{\sum w_{ij} p_{cij}}{w_{max}},$$

where p_{oij} is the proportion of the judgments observed in the ijth cell, p_{cij} is the proportion in the ijth cell expected by chance, and w_{ij} is the disagreement weight.

Table 16 shows the quadratic-weighted kappa for the training sample and the validation sample. The weighted kappa ranges from 0 to 1, where values of 0 indicate no agreement and values of 1 indicate perfect agreement. In the training sample, weighted kappa coefficients for operational writing prompts by dimension range from 0.52 to 0.82. In the validation sample, the range is from 0.55 to 0.79. The validation sample generally has higher or similarly weighted kappa compared to the training sample.

Table 16: Weighted Kappa Coefficients

Grade	Item ID	Dimension	Quadratic-Weighted Kappa	
			Two Human Raters	***Human and Machine***
3	7402	Convention	0.60	0.70
		Elaboration	0.60	0.69
		Purpose	0.67	0.69
	7407	Convention	0.65	0.69
		Elaboration	0.61	0.61

Grade	Item ID	Dimension	Quadratic-Weighted Kappa	
			Two Human Raters	***Human and Machine***
		Purpose	0.67	0.62
4	3084	Convention	0.64	0.68
		Elaboration	0.52	0.66
		Purpose	0.57	0.64
	3086	Convention	0.62	0.63
		Elaboration	0.58	0.55
		Purpose	0.60	0.61
5	3133	Convention	0.63	0.70
		Elaboration	0.54	0.62
		Purpose	0.65	0.66
	4286	Convention	0.65	0.65
		Elaboration	0.53	0.64
		Purpose	0.62	0.62
6	3138	Convention	0.55	0.65
		Elaboration	0.61	0.68
		Purpose	0.62	0.70
	5438	Convention	0.56	0.67
		Elaboration	0.58	0.70
		Purpose	0.62	0.71
7	3037	Convention	0.65	0.71
		Elaboration	0.61	0.67
		Purpose	0.59	0.63
	3883	Convention	0.60	0.66
		Elaboration	0.67	0.74
		Purpose	0.62	0.72
8	3056	Convention	0.69	0.73
		Elaboration	0.75	0.73
		Purpose	0.82	0.73
	3058	Convention	0.62	0.72
		Elaboration	0.75	0.72
		Purpose	0.72	0.79

The AIR AutoScoring Model can generate condition codes to indicate that the response provided by the student is considered invalid and therefore incorrect. All condition codes receive the lowest possible dimension score for purposes of ability estimation. The machine-generated condition codes, also referred to as rule-based condition codes, are as follows:

- NO_RESPONSE: No non-blank characters are detected in the response.
- NOT_ENOUGH_DATA: Student response is fewer than the minimum number of words configured in the rubric.
- PROMPT_COPY_MATCH: Student response is copied from the passage or item prompt (currently flagged when a 70% match is found, but this parameter is configurable).
- DUPLICATE_TEXT: Student response is repeated text copied over and over (currently flagged when a 70% match is found, but this parameter is configurable).
- NONSPECIFIC: Essay scoring engine predicts the assignment of a condition code. Even after training the system, there can be responses that do not fall into any of the pre-set categories. For those responses, the system will generate a condition code of NONSPECIFIC.

Based on AIRCore writing items administered, a confidence index is produced for each dimension of the prompts used for the spring 2019 writing assessment. To ensure the quality of the AutoScoring Model, responses that fall into one of these three scenarios were sent to AIR's Ohio Scoring Center to be scored by human readers: 1) the first 500 responses; 2) responses that received the lowest 15% of confidence index values; 3) any response that receives a condition code of NONSPECIFIC from the AutoScoring Model.

The human verification process was conducted by the sequence described below:

- If the verification reader assigned a score that was the same as the machine-assigned score, the machine-assigned score was accepted to be the final dimension score.

- If the first verification reader did not assign the same score as the machine-assigned score, the essay was sent to the second verification reader. If the second reader's score matched with either machine or the first reader's score, the matching score was accepted to be the final score.

- If the second verification reader's score did not match with the machine or first reader's score, the essay was sent to the scoring supervisor for assigning the final score.

- If a verification reader assigned a condition code, the condition code was accepted to be the final score.

Table 17 provides the agreement rate and quadratic-weighted Kappa coefficients between the scores provided by the AIR AutoScoring model and the first human verification reader in a sample of students who submitted their essay responses during an early period of the testing window.

Table 17: The First 500 Cases Percentage Agreement in Human-Scoring and AutoScoring

Grade	Item ID	Dimension	Human and AIR AutoScoring Model Agreement in the First 500 Cases				
			% Exact	% Adjacent	% Non-Adjacent	Q W Kappa	N
3	7402	Conventions	59.80	40.00	0.20	0.54	495
		Evidence and Elaboration	58.99	39.60	1.41	0.61	495
		Purpose, Focus, and Organization	56.16	41.41	2.42	0.62	495
	7407	Conventions	68.28	30.30	1.42	0.63	495
		Evidence and Elaboration	64.24	35.35	0.40	0.54	495
		Purpose, Focus, and Organization	72.12	27.47	0.40	0.59	495
4	3084	Conventions	61.70	38.10	0.20	0.55	496
		Evidence and Elaboration	76.41	22.98	0.60	0.52	496
		Purpose, Focus, and Organization	73.59	25.60	0.81	0.54	496
	3086	Conventions	66.20	33.60	0.20	0.51	497
		Evidence and Elaboration	61.77	35.41	2.82	0.45	497
		Purpose, Focus, and Organization	62.98	35.61	1.41	0.50	497
5	3133	Conventions	77.96	22.04	0.00	0.66	499
		Evidence and Elaboration	62.73	35.27	2.00	0.52	499
		Purpose, Focus, and Organization	54.31	41.68	4.01	0.47	499
	4286	Conventions	72.34	27.26	0.40	0.63	499
		Evidence and Elaboration	62.53	35.67	1.80	0.50	499
		Purpose, Focus, and Organization	67.33	31.86	0.80	0.55	499
6	3138	Conventions	76.36	23.64	0.00	0.65	499
		Evidence and Elaboration	56.71	41.88	1.40	0.51	499
		Purpose, Focus, and Organization	64.73	34.27	1.00	0.64	499
	5438	Conventions	70.30	28.88	0.80	0.59	495
		Evidence and Elaboration	64.24	35.15	0.61	0.58	495
		Purpose, Focus, and Organization	74.14	25.45	0.40	0.71	495

Grade	Item ID	Dimension	Human and AIR AutoScoring Model Agreement in the First 500 Cases				
			% Exact	% Adjacent	% Non-Adjacent	Q W Kappa	N
7	3037	Conventions	71.36	28.42	0.22	0.56	468
		Evidence and Elaboration	72.01	26.71	1.28	0.65	468
		Purpose, Focus, and Organization	67.52	32.48	0.00	0.59	468
	3883	Conventions	85.56	13.76	0.68	0.72	443
		Evidence and Elaboration	53.05	43.79	3.16	0.48	443
		Purpose, Focus, and Organization	59.37	39.73	0.90	0.53	443
8	3056	Conventions	83.30	16.70	0.00	0.68	485
		Evidence and Elaboration	70.52	29.07	0.41	0.74	485
		Purpose, Focus, and Organization	66.80	32.58	0.62	0.73	485
	3058	Conventions	83.00	16.40	0.60	0.63	494
		Evidence and Elaboration	62.96	36.84	0.20	0.63	494
		Purpose, Focus, and Organization	71.86	27.73	0.40	0.74	494

4. EVIDENCE OF CONTENT VALIDITY

This section demonstrates that the knowledge and skills assessed by the NH SAS were representative of the content standards of the larger knowledge domain. The content standards for NH SAS and the test development process are discussed, mapping NH SAS tests to the standards. A complete description of the test development process can be found in Volume 2. Further evidence of content validity will be provided in the future through a planned independent alignment study.

4.1 CONTENT STANDARDS

The NH SAS was aligned to the NH CCRS. The ELA and mathematics standards are available for review at http://www.education.nh.gov/instruction/curriculum/index.htm, and the science standards are available at http://www.education.nh.gov/instruction/curriculum/science/index.htm.

Table 18 through Table 20 present the reporting categories by grade and test, as well as the number of items administered measuring each category.

Table 18: Number of Items for Each Reporting Category, ELA

Reporting Category	Grade					
	3	***4***	***5***	***6***	***7***	***8***
Reading Informational Text (RI)	118	148	117	198	183	197
Reading Literary Text (RL)	106	108	84	120	149	95

**Note: Writing is not reported.*

Table 19: Number of Items for Each Reporting Category, Mathematics

Grade	Reporting Category	Number of Items
3	Measurement, Data, and Geometry (MDG)	84
	Numbers and Operations in Base Ten and Fractions (NBTF)	244
	Operations and Algebraic Thinking (OA)	151
4	Measurement, Data, and Geometry (MDG)	97
	Numbers and Operations in Base Ten and Fractions (NBTF)	301
	Operations and Algebraic Thinking (OA)	97
5	Measurement, Data, and Geometry (MDG)	81
	Numbers and Operations in Base Ten and Fractions (NBTF)	257
	Operations and Algebraic Thinking (OA)	71
6	Expressions and Equations (EE)	154
	Geometry, Statistics, and Probability (GSP)	71
	Ratios, Proportional Relationships, and the Number System (RPNS)	243
7	Expressions and Equations (EE)	65
	Geometry (G)	74
	Ratios, Proportional Relationships, and the Number System (RPNS)	134
	Statistics and Probability (SP)	65
8	Expressions, Equations, and the Number System (EENS)	161
	Functions (F)	87
	Geometry, Statistics, and Probability (GSP)	163

Table 20: Number of Items for Each Reporting Category, Science

Grade	Reporting Category	Cluster	Standalone
5	Earth and Space Science (ESS)	11	8
	Life Science (LS)	11	10
	Physics Science (PS)	8	11
8	Earth and Space Science (ESS)	8	7
	Life Science (LS)	6	12
	Physics Science (PS)	11	7
11	Earth and Space Science (ESS)	6	11
	Life Science (LS)	14	8
	Physics Science (PS)	8	8

5. Evidence on Internal-External Structure

In this section, the internal structure of the assessment is explored using the scores provided at the reporting-category level. The relationship of the subscores is just one indicator of the test dimensionality.

Scale scores and relative strengths and weaknesses based on each reporting category were provided to students. Evidence is needed to verify that scale scores and relative strengths and weaknesses for each reporting category provide both different and useful information for student performance.

It may not be reasonable to expect that the reporting category scores are completely orthogonal—this would suggest that there are no relationships among reporting category scores and would make justification of a unidimensional IRT model difficult, though reporting these separate scores could then easily be justified. On the contrary, if the reporting categories were perfectly correlated, a unidimensional model could be justified, but the reporting of separate scores could not.

One pathway to explore the internal structure of the test is via a second-order factor model, assuming a general mathematics construct (first factor) with reporting categories (second factor), and that the items load onto the reporting category they intend to measure. If the first-order factors are highly correlated, and the model fits data well for the second-order model, this provides evidence of unidimensionality, as well as of reporting subscores.

The science assessment is modeled with the Rasch testlet model (Wang & Wilson, 2005). Unlike the models for ELA and mathematics, the IRT model for science is a high-dimensional model, incorporating a nuisance dimension for each item cluster, in addition to an overall dimension representing the overall proficiency in science. This approach is innovative and quite different from the traditional approach of ignoring local dependencies. Validity evidence on the internal structure will focus on the presence of cluster effects and how substantial they are.

Another pathway is to explore observed correlations between the subscores. However, as each reporting category is measured with a small number of items, the standard errors of the observed scores within each reporting category are typically larger than the standard error of the total test score. Disattenuating for measurement error could offer some insight into the theoretical true score correlations. Both observed correlations and disattenuated correlations are provided in the following section.

5.1 Correlations Among Reporting Category Scores

The correlations among reporting category scores, both observed (below diagonal) and corrected for attenuation (above diagonal) are presented in Table 21 through Table 23. On the diagonal, the reliability coefficient of the reporting category is shown. In ELA, the observed correlations among the reporting categories range from 0.56 to 0.66. For mathematics, the observed correlations were between 0.51 and 0.79. For science, the observed correlations were between 0.56 and 0.61. Disattenuated correlations were between 0.75 and 0.87 for ELA, 0.70 and 0.96 for mathematics, and 0.85 and 0.92 for science.

In some instances, these correlations were lower than one might expect. However, as previously noted, the correlations were subject to a large amount of measurement error at the strand level, given the limited number of items from which the scores were derived. Consequently, over-interpretation of these correlations, as either high or low, should be made cautiously.

Table 21: Correlations Among Reporting Categories, ELA

Grade	Reporting Category	Mean # of Items Per Student	Cat1	Cat2
3	Reading Informational Text (Cat1)	15.4	0.74*	0.75
	Reading Literary Text (Cat2)	15.6	0.56	0.76*
4	Reading Informational Text (Cat1)	15.8	0.74*	0.86
	Reading Literary Text (Cat2)	15.6	0.64	0.75*
5	Reading Informational Text (Cat1)	15.2	0.73*	0.87
	Reading Literary Text (Cat2)	15.7	0.66	0.79*
6	Reading Informational Text (Cat1)	15.6	0.75*	0.84
	Reading Literary Text (Cat2)	15.4	0.62	0.73*
7	Reading Informational Text (Cat1)	15.4	0.76*	0.85
	Reading Literary Text (Cat2)	15.4	0.64	0.75*
8	Reading Informational Text (Cat1)	15.4	0.76*	0.86
	Reading Literary Text (Cat2)	15.7	0.65	0.75*

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal and disattenuated are above.*

Table 22: Correlations Among Reporting Categories, Mathematics

Grade	Reporting Category	Mean # of Items Per Student	Cat1	Cat2	Cat3	Cat4
3	Measurement, Data, and Geometry (Cat1)	9.0	0.76*	0.92	0.87	-
	Numbers and Operations in Base Ten and Fractions (Cat2)	14.0	0.73	0.83*	0.93	-
	Operations and Algebraic Thinking (Cat3)	10.9	0.68	0.76	0.81*	-
4	Measurement, Data, and Geometry (Cat1)	9.0	0.73*	0.91	0.87	-
	Numbers and Operations in Base Ten and Fractions (Cat2)	15.9	0.72	0.86*	0.96	-
	Operations and Algebraic Thinking (Cat3)	8.9	0.65	0.78	0.77*	-
5	Measurement, Data, and Geometry (Cat1)	10.1	0.76*	0.91	0.82	-
	Numbers and Operations in Base Ten and Fractions (Cat2)	15.6	0.73	0.85*	0.88	-
	Operations and Algebraic Thinking (Cat3)	8.3	0.62	0.70	0.75*	-
6	Expressions and Equations (Cat1)	11.8	0.78*	0.70	0.94	-
	Geometry, Statistics, and Probability (Cat2)	8.0	0.51	0.69*	0.70	-
	Ratios, Proportional Relationships, and the Number System (Cat3)	14.2	0.76	0.53	0.84*	-
7	Expressions and Equations (Cat1)	8.8	0.75*	0.83	0.89	0.83
	Geometry (Cat2)	8.0	0.61	0.72*	0.86	0.79
	Ratios, Proportional Relationships, and the Number System (Cat3)	8.5	0.69	0.65	0.80*	0.88
	Statistics and Probability (Cat4)	8.7	0.61	0.57	0.67	0.72*
8	Expressions, Equations, and the Number System (Cat1)	11.0	0.83*	0.93	0.95	-
	Functions (Cat2)	8.9	0.73	0.74*	0.91	-
	Geometry, Statistics, and Probability (Cat3)	13.9	0.79	0.72	0.84*	-

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal and disattenuated are above.*

Table 23: Correlations Among Reporting Categories & Reporting Category Reliabilities, Science

Grade	Reporting Category	# of Items	Earth and Space Science	Life Science	Physical Science
5	Earth and Space Science		0.69*	0.88	0.88
	Life Science		0.58	0.63*	0.92
	Physical Science		0.57	0.57	0.61*
8	Earth and Space Science		0.66*	0.90	0.89
	Life Science		0.61	0.69*	0.88
	Physical Science		0.57	0.58	0.63*
11	Earth and Space Science		0.66*	0.85	0.87
	Life Science		0.59	0.72*	0.85
	Physical Science		0.56	0.58	0.63*

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal and disattenuated are above.*

5.2 Convergent and Discriminant Validity

According to Standard 1.16 of the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), evidence must be provided of convergent and discriminant validity, a part of validity evidence demonstrating that assessment scores are related as expected with criterion and other variables for all student groups. However, a second, independent test measuring the same constructs as mathematics and ELA in New Hampshire during the same time period, which could easily permit for a cross test set of correlations, was not available. Therefore, as an alternative, the correlations between subscores within and across mathematics and ELA were examined. The *a priori* expectation is that subscores within the same subject (e.g., mathematics) will correlate more positively than subscore correlations across subjects (e.g., mathematics and ELA). These correlations are based on a small number of items; consequently, the observed score correlations will be smaller in magnitude as a result of the very large measurement error at the subscore level. For this reason, both the observed correlations and the disattenuated correlations are provided.

Observed and disattenuated subscore correlations were calculated both within and across subjects for grades 3–8 mathematics and ELA. Generally, the pattern is consistent with the a priori expectation that subscores within a test correlate more highly than correlations between tests measuring a different construct. The correlations among reporting category scores, both observed (below diagonal) and corrected for attenuation (above diagonal) are presented in Table 21 through

Table 29. On the diagonal, the reliability coefficient of the reporting category is shown.

Table 24: Correlations Across Subjects, Grade 3

Subject	Number of Students	Reporting Category	ELA		Mathematics		
			Cat1	Cat2	Cat1	Cat2	Cat3
ELA	11,183	Reading Informational Text (Cat1)	0.74*	0.75	0.67	0.68	0.67
		Reading Literary Text (Cat2)	0.56	0.76*	0.66	0.68	0.66
Mathematics		Measurement, Data, and Geometry (Cat1)	0.50	0.50	0.76*	0.92	0.87
		Numbers and Operations in Base Ten & Fractions (Cat2)	0.53	0.54	0.73	0.83*	0.91
		Operations and Algebraic Thinking (Cat3)	0.52	0.52	0.68	0.75	0.81*

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal and disattenuated are above.*

Table 25: Correlations Across Subjects, Grade 4

Subject	Number of Students	Reporting Category	ELA		Mathematics		
			Cat1	Cat2	Cat1	Cat2	Cat3
ELA	11,606	Reading Informational Text (Cat1)	0.74*	0.86	0.71	0.73	0.73
		Reading Literary Text (Cat2)	0.64	0.75*	0.70	0.72	0.74
Mathematics		Measurement, Data, and Geometry (Cat1)	0.52	0.52	0.73*	0.91	0.87
		Numbers and Operations in Base Ten & Fractions (Cat2)	0.58	0.58	0.72	0.86*	0.96
		Operations and Algebraic Thinking (Cat3)	0.55	0.56	0.65	0.78	0.77*

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal and disattenuated are above.*

Table 26: Correlations Across Subjects, Grade 5

Subject	Number of Students	Reporting Category	ELA		Mathematics			Science		
			Cat1	Cat2	Cat1	Cat2	Cat3	Cat1	Cat2	Cat3
ELA	11,764	Reading Informational Text (Cat1)	0.73*	0.87	0.70	0.76	0.75	0.79	0.85	0.83
		Reading Literary Text (Cat2)	0.66	0.79*	0.69	0.73	0.73	0.76	0.84	0.80
Mathematics		Measurement, Data, and Geometry (Cat1)	0.53	0.53	0.76*	0.90	0.82	0.74	0.74	0.79
		Numbers and Operations in Base Ten & Fractions (Cat2)	0.60	0.59	0.72	0.85*	0.87	0.78	0.78	0.81
		Operations and Algebraic Thinking (Cat3)	0.56	0.56	0.62	0.69	0.75*	0.75	0.76	0.79
Science		Earth and Space Science (Cat1)	0.56	0.56	0.54	0.59	0.54	0.69*	0.88	0.88
		Life Science (Cat2)	0.58	0.59	0.51	0.57	0.52	0.58	0.63*	0.92
		Physical Science (Cat3)	0.55	0.55	0.54	0.58	0.54	0.57	0.57	0.61*

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal and disattenuated are above.*

Table 27: Correlations Across Subjects, Grade 6

Subject	Number of Students	Reporting Category	ELA		Mathematics		
			Cat1	Cat2	Cat1	Cat2	Cat3
ELA	12,246	Reading Informational Text (Cat1)	0.75*	0.84	0.72	0.57	0.73
		Reading Literary Text (Cat2)	0.62	0.73*	0.74	0.58	0.75
Mathematics		Expressions and Equations (Cat1)	0.55	0.56	0.78*	0.68	0.93
		Geometry & Statistics and Probability (Cat2)	0.41	0.41	0.50	0.69*	0.70
		Ratios and Proportional Relationships & Number System (Cat3)	0.58	0.59	0.75	0.53	0.84*

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal and disattenuated are above.*

Table 28: Correlations Across Subjects, Grade 7

Subject	Number of Students	Reporting Category	ELA		Mathematics			
			Cat1	Cat2	Cat1	Cat2	Cat3	Cat4
ELA	12,146	Reading Informational Text (Cat1)	0.76*	0.85	0.75	0.70	0.76	0.74
		Reading Literary Text (Cat2)	0.64	0.75*	0.75	0.69	0.74	0.73
Mathematics		Expressions and Equations (Cat1)	0.57	0.56	0.75*	0.83	0.89	0.83
		Geometry (Cat2)	0.52	0.51	0.61	0.72*	0.86	0.79
		Ratios and Proportional Relationships & Number System (Cat3)	0.59	0.57	0.69	0.65	0.80*	0.88
		Statistics and Probability (Cat4)	0.55	0.54	0.61	0.57	0.67	0.72*

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal and disattenuated are above.*

Table 29: Correlations Across Subjects, Grade 8

Subject	Number of Students	Reporting Category	ELA		Mathematics			Science		
			Cat1	Cat2	Cat1	Cat2	Cat3	Cat1	Cat2	Cat3
ELA	11,890	Reading Informational Text (Cat1)	0.76*	0.84	0.73	0.74	0.73	0.8	0.79	0.78
		Reading Literary Text (Cat2)	0.63	0.75*	0.72	0.71	0.72	0.79	0.78	0.76
Mathematics		Expressions and Equations & Number System (Cat1)	0.58	0.56	0.83*	0.92	0.95	0.82	0.81	0.82
		Functions (Cat2)	0.55	0.52	0.72	0.74*	0.91	0.82	0.8	0.81
		Geometry & Statistics and Probability (Cat3)	0.58	0.57	0.79	0.71	0.84*	0.82	0.81	0.83
Science		Earth and Space Science (Cat1)	0.56	0.55	0.61	0.57	0.61	0.66*	0.9	0.89
		Life Science (Cat2)	0.57	0.56	0.61	0.57	0.62	0.61	0.69*	0.88
		Physical Science (Cat3)	0.54	0.52	0.59	0.55	0.6	0.57	0.58	0.63*

**Diagonal value represents the reliability coefficient of the reporting category. Observed correlations are below the diagonal and disattenuated are above.*

Additionally, the correlation was computed among the overall scores for the three tested subjects: ELA, mathematics, and science. Correlations are presented in Table 30 and are relatively high, between 0.74 and 0.77.

Table 30: Correlations Across Spring 2019 ELA, Mathematics, and Science Scores

Grade	N	ELA & Mathematics	ELA & Science	Mathematics & Science
5	11,781	0.74	0.76	0.74
8	11,916	0.75	0.76	0.77

5.2.1 Summative and Interim Correlations

Beginning in fall 2018 and continuing through spring 2019, optional ELA and mathematics interim assessments were administered. These tests were online and adaptive. Test takers who took both the summative assessment in spring 2019 and optional interim assessments were identified for conducting the cross-test set of correlations. Table 31 and Table 32 present the correlations between summative and interim assessments for ELA and mathematics. Observed correlations are medium to high, ranging from 0.70 to 0.85. Disattenuated correlations are relatively higher, with a range from 0.80 to 0.97. The number (N) of students, mean, and standard deviation of scale score, and reliability coefficient reported in tables are based on students who took both the summative assessment and the interim assessment.

Table 31: Summative vs. Interim Correlations, ELA

Grade	Test	Scale Score Mean	Scale Score SD	Reliability Coefficient	Observed Correlation	Disattenuated Correlation	N
3	Summative	587.6	38.76	0.89	0.73	0.84	1,073
	Interim	579.62	41.69	0.84			
4	Summative	617.60	43.73	0.89	0.78	0.89	898
	Interim	602.36	46.42	0.86			
5	Summative	629.87	40.99	0.89	0.79	0.90	649
	Interim	623.85	46.66	0.86			
6	Summative	641.71	43.65	0.90	0.77	0.88	545
	Interim	630.4	45.72	0.86			
7	Summative	646.71	45.24	0.90	0.70	0.80	598
	Interim	633.72	43.70	0.85			
8	Summative	663.64	45.06	0.90	0.74	0.84	309
	Interim	643.43	53.90	0.86			

Table 32: Summative vs. Interim Correlations, Mathematics

Grade	Test	Scale Score Mean	Scale Score SD	Reliability Coefficient	Observed Correlation	Disattenuated Correlation	N
3	Summative	436.47	31.34	0.92	0.76	0.84	1,268
	Interim	418.61	33.26	0.88			
4	Summative	462.47	38.75	0.92	0.78	0.87	1,415
	Interim	447.95	43.20	0.88			
5	Summative	492.18	47.11	0.91	0.81	0.92	1,711
	Interim	472.79	48.95	0.85			
6	Summative	527.62	51.18	0.91	0.77	0.87	844
	Interim	502.11	55.80	0.87			
7	Summative	541.39	57.53	0.90	0.85	0.97	430
	Interim	529.78	58.49	0.86			
8	Summative	587.41	71.97	0.93	0.80	0.89	511
	Interim	563.94	75.40	0.86			

5.3 Relationship of Test Scores to External Variables

The relationship of test scores to external variables measuring the same or related constructs is an important source of validity evidence. The NH SAS was first administered to students during the spring of 2018, replacing SBAC in ELA and mathematics and the NECAP in science. Ideally, we would correlate two different tests measuring a common construct administered within a similar time period. Here, we present correlations between two different tests measuring a common construct but measured using the same students one year apart. We expect the correlations to be high to suggest that the NH SAS has a high relationship with an externally developed measure, though the time gap between the two different assessments is greater than if the two tests were measured within a similar testing window. Table 33 and Table 34 present correlations between SBAC scores from spring 2017 and NH SAS scores from spring 2018. Observed correlations are between 0.77 and 0.86, and disattenuated correlations are between 0.86 and 0.93, both of which can be considered relatively high compared to industry standards.

Table 33: Correlations Between Spring 2017 SBAC Scores and Spring 2018 NH SAS Scores, ELA

Grade in Spring 2017	Grade in Spring 2018	N	Observed Correlations	Disattenuated Correlations
3	4	11,173	0.77	0.86
4	5	11,219	0.80	0.89
5	6	11,381	0.81	0.90

Grade in Spring 2017	Grade in Spring 2018	N	Observed Correlations	Disattenuated Correlations
6	7	11,333	0.81	0.90
7	8	11,776	0.81	0.90

Table 34: Correlations Between Spring 2017 SBAC Scores and Spring 2018 NH SAS Scores, Mathematics

Grade in Spring 2017	Grade in Spring 2018	N	Observed Correlations	Disattenuated Correlations
3	4	11,479	0.80	0.86
4	5	11,328	0.82	0.89
5	6	11,365	0.82	0.89
6	7	11,356	0.86	0.93
7	8	11,764	0.85	0.91

5.4 CLUSTER EFFECTS FOR SCIENCE

The NH SAS for science uses the Rasch testlet model (Wang & Wilson, 2005). Unlike the models for ELA and mathematics, the IRT model for science is a high-dimensional model, incorporating a nuisance dimension for each item cluster, in addition to an overall dimension representing the overall proficiency in science. A detailed description of the IRT model, including an illustration using a directed graph in Figure 1, is shown in Volume 1, Section 5.2. The psychometric approach for the science assessment is innovative and quite different from the traditional approach of ignoring local dependencies. The validity evidence on the internal structure presented in this section relates to the presence of cluster effects and how substantial they are.

Simulation studies conducted by Rijmen, Jiang, and Turhan (2018) confirmed that both the item difficulty parameters and the cluster variances are recovered well for the Rasch testlet model under a variety of conditions. Cluster effects with a range of magnitudes were recovered well. The results obtained by Rijmen, Jiang, and Turhan (2018) confirmed earlier findings reported in the literature (e.g., Bradlow, Wainer, & Wang, 1999) under conditions that were chosen to closely resemble the science assessment. For example, in one of the studies, the item location parameters and cluster variances used to simulate data were based on the results of a pilot study.

We examined the distribution of cluster variances obtained from 2018 IRT calibration. For elementary school, the estimated value of the cluster variances of all operational, scored items ranged from 0 to 4.46, with a median value of 0.47 and a mean value of 0.81. The median value is slightly smaller than the estimated variance parameter of the overall science dimension ($\hat{\sigma}_\theta^2 = 0.61$). For middle school, the estimated value of the cluster variances of all operational, scored items ranged from 0.07 to 1.29, with a median value of 0.40 and a mean value of 0.46. The median value is close to the estimated variance parameter of the overall science dimension ($\hat{\sigma}_\theta^2 = 0.44$). For high school, the estimated value of cluster variances of all operational, scored items ranged from 0.10 to 0.95, with a median value of 0.40 and a mean value of 0.43. The median value is slightly smaller than the estimated variance parameter of the overall science dimension ($\hat{\sigma}_\theta^2 = 0.61$). Figure 5 through Figure 7 present the histograms of the cluster variances expressed as the proportion of the total variance for all operational items for elementary, middle, and high school, respectively. For all grade bands, a wide range of cluster variances is observed. These results indicate that, for both grades, cluster effects can be substantial and provide evidence for the appropriateness of a psychometric model that explicitly takes into account local dependencies among the assertions of an item cluster.

Figure 5: Cluster Variance Proportion for Science Operational Items in Elementary School

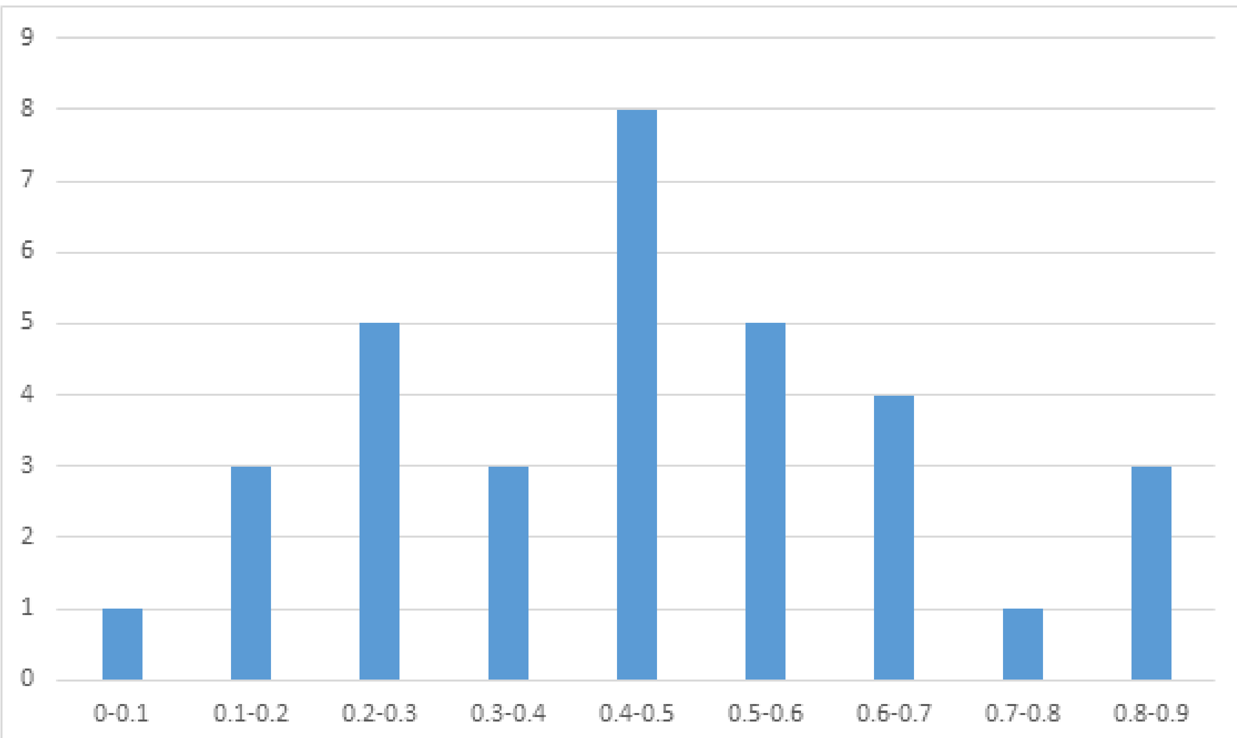


Figure 6: Cluster Variance Proportion for Science Operational Items in Middle School

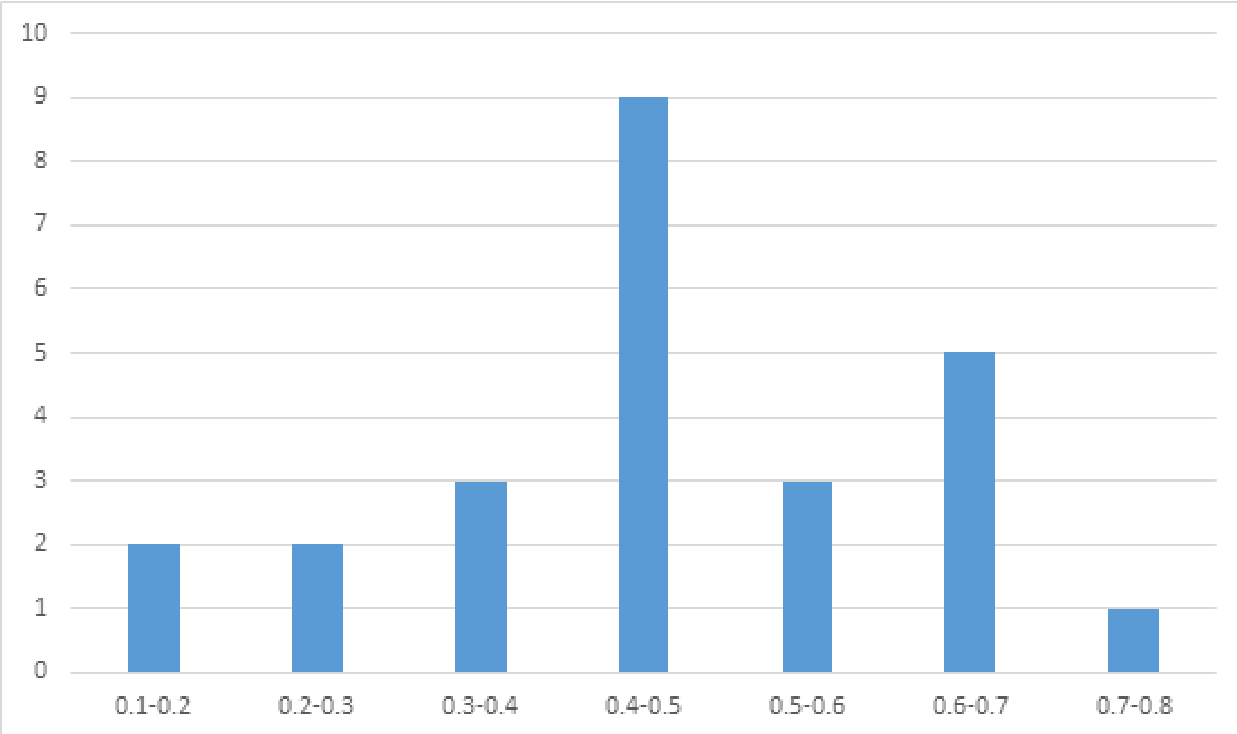
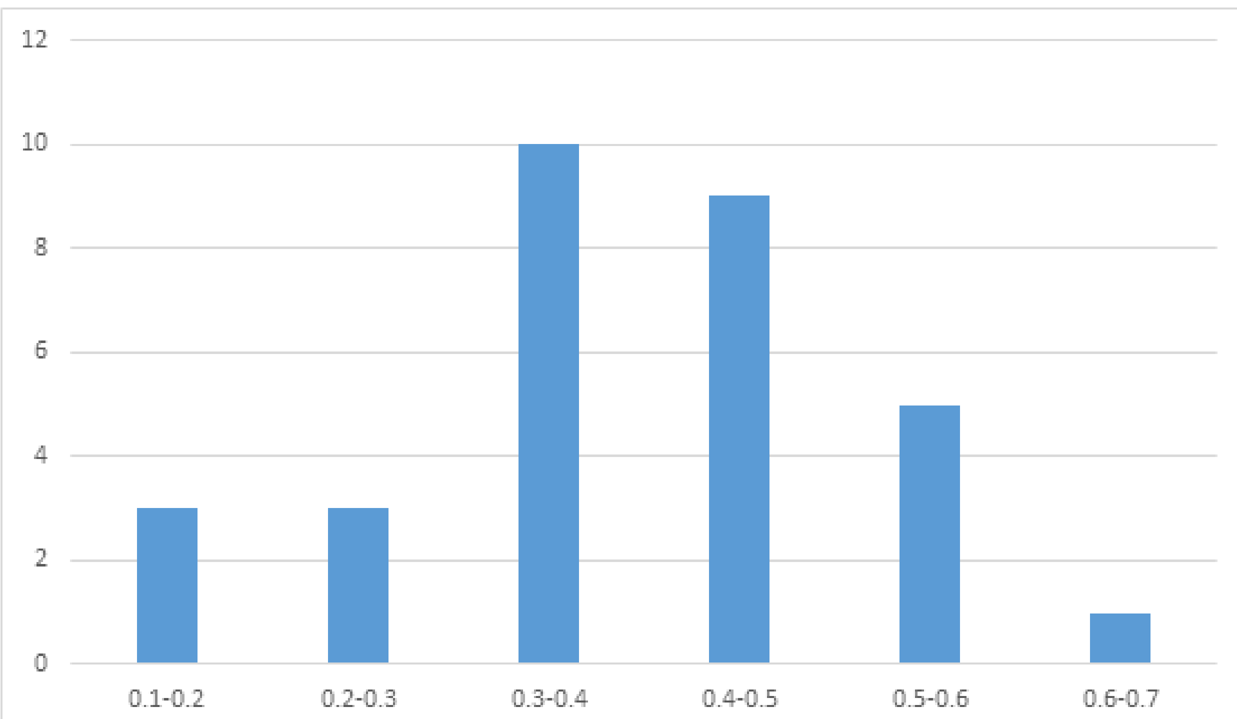


Figure 7: Cluster Variance Proportion for Science Operational Items in High School



6. Fairness and Accessibility

6.1 Fairness in Content

The principles of universal design of assessments provide guidelines for test design to minimize the impact of construct-irrelevant factors in assessing student achievement. Universal design removes barriers to provide access for the widest range of students possible. Seven principles of universal design are applied in the process of test development (Thompson, Johnstone, & Thurlow, 2002):

1. Inclusive assessment population
2. Precisely defined constructs
3. Accessible, non-biased items
4. Amenable to accommodations
5. Simple, clear, and intuitive instructions and procedures
6. Maximum readability and comprehensibility
7. Maximum legibility

Test development specialists have received extensive training on the principles of universal design and apply these principles in the development of all test materials. In the review process, adherence to the principles of universal design is verified by New Hampshire educators and stakeholders.

6.2 Statistical Fairness in ELA and Mathematics Item Statistics

Due to the use of adaptive testing in the NH SAS for ELA and mathematics, the number of New Hampshire students who see each item is relatively small. DIF analysis for the NH SAS for ELA and mathematics is not available due to the small sample size for each demographic group. However, DIF analysis was conducted with other states that field tested the items. A thorough content review was performed in those states. The details surrounding this review of items for bias is further described in Volume 1, Section 4.4.

7. Summary

This report is intended to provide a collection of reliability and validity evidence to support appropriate inferences from the observed test scores. The overall results can be summarized as follows:

- Reliability: Reliability estimates are provided at the aggregate and subgroup levels, showing that the reliability of all tests is in line with acceptable industry standards.
- Content validity: Evidence is provided to support the assertion that content coverage on each form was consistent with test specifications of the blueprint across testing modes.
- Internal structural validity: Evidence is provided to support the reporting of an overall score and subscores at the reporting category levels.
- Relationship of test scores to external variables: Evidence of convergent and discriminant validity is provided to support the relationship between the test and other measures intended to assess similar constructs, as well as the relationship between the test from other measures intended to assess different constructs.

8. References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: Author.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153–168.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment, Research & Evaluation, 11*(6).

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13–103). New York: Macmillan.

Rijmen, F., Jiang, T, & Turhan, A. (2018, April). An Item Response Theory Model for New Science Assessments. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, *7*(14).

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 2002, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126–149.